Providing Automatic Multilingual Text Generation to Artificial Cognitive Systems

Carles Fernández
Computer Vision Centre
Universitat Autònoma de Barcelona, Spain

Xavier Roca
Computer Vision Centre
Universitat Autònoma de Barcelona, Spain

Jordi Gonzàlez
Institut de Robòtica i Informàtica Ind.
Universitat Politècnica de Catalunya,Spain

{perno | xavir | poal}@cvc.uab.es

Abstract

This contribution addresses the incorporation of a module for advanced user interaction into an artificial cognitive vision system to include the *human-in-the-loop*. Specifically, the document describes a method to automatically generate natural language textual descriptions of meaningful events and behaviors, in a controlled scenario. One of the goals of the system is to be capable of producing these descriptions in multiple languages. We will introduce some relevant stages of the whole system, and concentrate on the linguistic aspects which have been taken into account to derive final text from conceptual predicates. Some experimental results are provided for the description of simple and complex behaviors of pedestrians in an intercity crosswalk, for Catalan, English, Italian, and Spanish.

Keywords: natural language generation, behavior analysis, multilingual generation, human sequence evaluation, artificial cognitive system.

Resumen

Esta contribución trata de la incorporación de un módulo de interacción avanzada entre usuarios en un sistema de visión cognitiva artificial para incluir el "human-in-the loop". El documento describe un método para generar, automáticamente, descripciones textuales de lenguaje natural de sucesos y comportamientos que tienen sentido, en un ambiente controlado. Una de las metas del sistema es la capacidad de producir estas descripciones en lenguajes múltiples. Introduciremos algunas fases relevantes del sistema completo, concentrándonos en los aspectos lingüísticos que se han tenido en cuenta para derivar el texto final a partir de predicados conceptuales. Se presentan algunos resultados experimentales relacionados con la descripción de comportamientos simples y complejos de viandantes en un paso de zebra de una ciudad, para las lenguas inglés, italiano y español.

Palabras clave: generación de lenguaje natural, análisis de comportamiento, generación multilingüe, evaluación de secuencias humanas, sistema cognitivo artificial.

1. Introduction

The introduction of Natural Language (NL) interfaces into vision systems has become popular, especially for surveillance systems (Gerber & Nagel, 2008). In this kind of applications, human behavior is represented by predefined sequences of events. Scenes are evaluated and automatically translated into text by analyzing the contents of the images over time, and deciding on the most suitable predefined event that applies in each case.

Such a process is referred to as Human Sequence Evaluation (HSE) in Gonzàlez, Rowe, Varona, & Roca, 2008. HSE takes advantage of cognitive capabilities for the semantic understanding of observed situations involving persons. This conception aims to perform an automatic evaluation of generally complex human behavior from image sequences in restricted discourse domains. In our case, the domain of interest has been restricted to urban outdoor surveillance environments.

This automatic analysis and description of temporal events was already tackled by Marburger et al. (Marburger, Neumann, & Novak, 1981), who proposed a NL dialogue system in German to retrieve information about traffic scenes. More recent methods for describing human activities from video images have been reported by Kojima et al. (Kojima et al., 2000; Kojima, Tamura, & Fukunaga, 2002), and automatic visual surveillance systems for traffic applications have been studied in Nagel, 2004 and Buxton & Gong, 1995, among others. These approaches present one or more specific limitations such as textual generation in a single language, surveillance for vehicular traffic applications only, restrictions for uncertain data, or very rigid environments.

There exist interesting approaches in some of the specific tasks presented here. Hovy, 1993 describes work done in discourse generation using discourse structure relations, especially regarding automated planning and generation of text containing multiple sentences. Emele et al., 1990 propose an architecture for the organization of linguistic knowledge for multilingual generation, based on typed feature structures. More recently, Lou et al., 2002 discuss a general framework for semantic interpretation of vehicle and pedestrian behaviors in visual traffic surveillance scenes.

We aim to build a system that addresses the aforementioned limitations, viz. monolingual generation, exclusivity of the application domain, uncertainty management, and rigidness, by following the proposals of HSE, in order to generate NL descriptions of human behavior appearing in controlled scenarios. There exist several considerations that have been taken into account for the design of such a system towards this goal:

- The resulting system should be *flexible* enough to: (i) enable a multilingual generation of discourse in natural language with average external users, and (ii) enable such a discourse to address the communication of complex events happening in the observed scenario, e.g. interactions among entities from more than one application domain (surveillance over both pedestrians and vehicles), contextualization of actions in a metric-temporal framework, or statements about reasoned interpretations for certain situations.
- This system has also been *restricted* to cover a defined domain of interest, given by the tackled outdoor inner city scenario and the model of possible situations to expect. As a result, we work with particularized linguistic models, which however must still be able to automatically produce natural descriptions of the occurring facts.

Experimental results have been focused to be specialized to a single type of scenario in order to study the problems in-depth, rather than attempting to come up with a supposedly generally applicable solution. This agrees with the *situatedness* property of cognitive systems (Wilson & Keil, 2001). Two particular scenes have been considered, which contain complex situations resulting from the interaction of pedestrians and vehicles in an outdoor environment, see **Figure 1**. Both consist of crosswalk scenes, in which pedestrians, cars, and objects appear and interact. On the first scene, four pedestrians cross the road in different ways. Several behaviors appear on the second one, e.g. displacements, meetings, crossings, accelerations, object disposals, and more complex situations such as abandoned objects, dangers of running over, and thefts. Hence, we consider the first scene as simpler than the second one, in terms of the complexity of the behaviors and interactions appearing, and the semantic analysis required to extract interpretations from them. The recording has been obtained using a distributed system of static cameras, and the scenario has been modeled a priori.

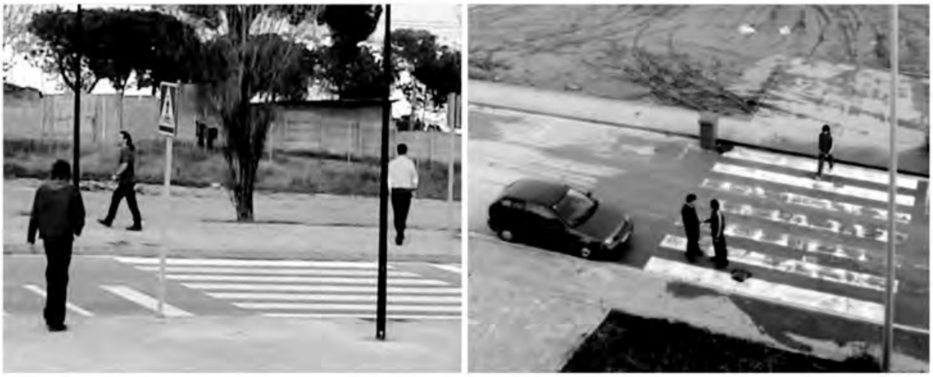


Figure 1: *Left: crosswalk scene showing simple behavior. Right: crosswalk scene including some more complex behaviors and interactions.*

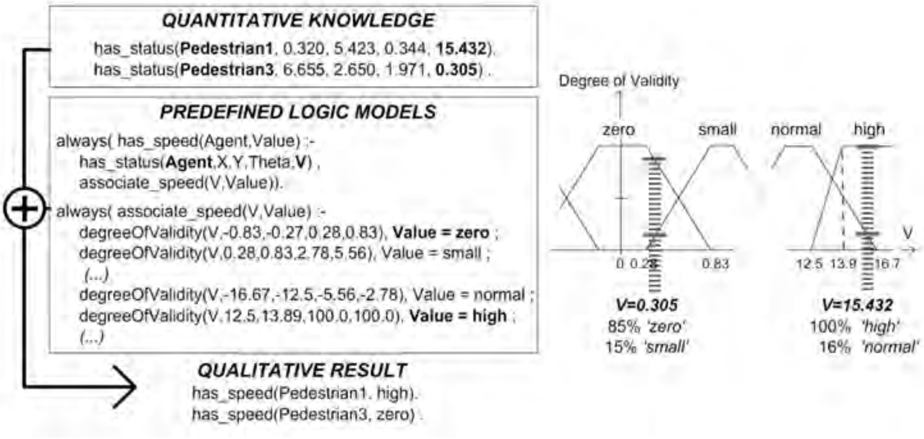
Next section provides a brief overview about the results obtained at the vision and conceptual levels. After that, we detail the main stages and tasks accomplished specifically at the NL Generation (NLG) module. Finally, some results are shown and evaluated, and last section highlights some general ideas and concludes the work.

2. Vision and Conceptual Levels

The *Vision level* acquires relevant visual content from the scenes by using a distribution of cameras. The detection and capture of interesting objects within the images is accomplished at this stage, by means of segmentation and tracking procedures which capture the motion information (Huerta et al., 2007; Rowe et al., 2005). As a result, a series of quantitative measurements over time is provided for each detected target, such as positions, velocities, and orientations of the agents.

In our case, we distinguish among different concepts within a scene, namely *agents*, being pedestrians and vehicles; *objects*, for movable objects like bags or relevant static elements of the scenario, e.g., benches; *locations* for interesting areas of the scenario, such as sidewalks, crosswalks, or waiting regions; and *events* for the actions, situations, or behaviors expected in a domain of interest.

Although we could express the observed facts in a quantitative way, e.g., "*The vehicle moved to the right at a speed of 23 km/h*", natural language is more inclined to be vague and inexact, and to use fuzzy prototypical concepts in order to evaluate facts in linguistic terms. Then, it would be better to say that the vehicle is moving at *low*, *medium*, or *high* speed depending on the context of this observation, also to deal with the inherent uncertainty of the assertions, and to better relate and categorize the situations we observe. The *conceptual level* accomplishes the conversion from quantitative to qualitative information. First, spatiotemporal data is represented by means of logical predicates created for each frame of the video sequence, in which numerical information is represented by its membership to predefined fuzzy functions. For example, a zero, small, average or high tag can be assigned, depending on the instantaneous velocity value (V) for an agent, see **Figure 2.** Apart from categorizing instantaneous facts, a scenario model also enables us to situate agents and objects in meaningful regions of the recorded location, e.g. crosswalk, sidewalk, or waiting zones.



```
has_status(Pedestrian1, 0.320, 5.423, 0.344, 15.432).
has_status(Pedestrian3, 6.655, 2.650, 1.971, 0.305) .

always( has_speed(Agent,Value) :-
   has_status(Agent,X,Y,Theta,V) ,
   associate_speed(V,Value)).
always( associate_speed(V,Value) :-
   degreeOfValidity(V,-0.83,-0.27,0.28,0.83), Value = zero ;
   degreeOfValidity(V,0.28,0.83,2.78,5.56), Value = small ;
   (...)
   degreeOfValidity(V,-16.67,-12.5,-5.56,-2.78), Value = normal ;
   degreeOfValidity(V,12.5,13.89,100.0,100.0), Value = high ;
   (...)

has_speed(Pedestrian1, high).
has_speed(Pedestrian3, zero) .
```

Figure 2 *– Conversion from quantitative to qualitative values. The numerical value of velocity for an agent (last field of* has_status*) at a time step is linked to the most probable membership of the* has_speed *fuzzy function.*

Nevertheless, we obtained a large collection of basic geometric facts, i.e., information about geometric properties such as positions and velocities, which needs to be filtered so that relevant information and patterns are extracted from it. Specifically, our aim is to detect admissible sequences of occurrences, which will contextualize geometric and temporal information about the scene, and will allow us to interpret the situation an agent is in. For instance, a sequence in which an agent walks by a sidewalk and stops in front of a crosswalk probably means that this agent is *waiting to cross*.

Situation Graph Trees are the specific tool used to build these models (Arens & Nagel, 2003; Gonzàlez, Rowe, Varona, & Roca, 2008), see **Figure 3**. They connect a set of defined situations by means of prediction and specialization edges. When a set of conditions is asserted, a high-level predicate is produced as an interpretation of a situation. An interesting property at this point is that the produced notes are much closer to a linguistic reading, since they interrelate and put into context different semantic elements such as locations, agents, and objects. Nevertheless, these expressions still keep language independence, and hence are a good starting point for multilingual text generation. More information about this situational analysis can be found in Fernández, Baiget, Roca, & Gonzàlez, 2007.

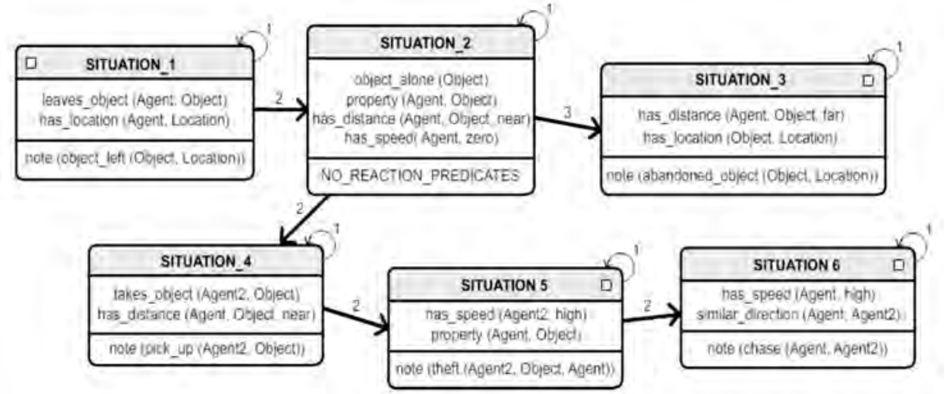

SITUATION_1
leaves_object (Agent. Object)
has_location (Agent, Location)
note (object_left (Object, Location))
SITUATION_2
object_alone (Object)
property (Agent, Object)
has_distance (Agent, Object near)
has_speed(Agent, zero)
NO_REACTION_PREDICATES
SITUATION_3
has_distance (Agent. Object. far)
has_location (Object. Location)
note (abandoned_object (Object, Location))
SITUATION_4
takes_object (Agent2, Object)
has_distance (Agent, Object near)
note (pick_up (Agent2. Object))
SITUATION 5
has_speed (Agent2 high)
property (Agent, Object)
note (theft (Agent2, Object, Agent))
SITUATION 6
has_speed (Agent, high)
similar_direction (Agent, Agent2)
note (chase (Agent, Agent2))
1
2
3

Figure 3 – *Situation Graph Trees are used to model situations and behaviors as predefined sequences of basic events. The example shown allows for complex inferences such as abandoned objects, chasings or thefts, by means of high-level* note *predicates.*

3. The NLG Module

NLG can be seen as a subfield of both computer science and cognitive science. It focuses on computer systems which can automatically produce understandable texts in a natural human language, so it is concerned with computational models of language and its use. NLG has been often considered as a process of *choice*, in which the most suitable mean has to be selected to achieve some desired end (Reiter & Dale, 2000).

The set of situations that need to be expressed are modeled and made available to the purposed NLG module, so that the main goal for this module consists of selecting one unique form of expressing that information in a clear and natural way, for each of the languages considered. This module is then built from a deterministic point of view, since it deals with aforeknown situational models.

Reiter & Dale, 2000 present a roadmap of the main tasks to be solved regarding NL text generation. Its proposed model of architecture includes three modules, see **Figure 4**:

- A *Document Planner*, which produces a specification of the text's content and structure, i.e. what has to be communicated by the NLG, by using both domain knowledge and practical information to be embedded into text.
- A *Microplanner*, in charge of filling the missing details regarding the concrete implementation document structure, i.e. in which way the information has to be communicated: distribution, referring expressions, level of detail, voice, etc.
- A *Surface Realizer*, which converts the abstract specification given by the previous stages into a real text, possibly embedded within some medium. It involves traversing the nodal text specification until the final presentation form.

Our described system is based on this generic approach, and enhances it by including multilingual capabilities and situation-guided content planning. Visual trackers acquire basic quantitative information about the scene, and the reasoning system decides how this information needs to be structured, gives coherency to the results, and also carries out inferences based on predefined conceptual and situational models. All these tasks are related to the Document Planner, since they provide the structured knowledge to be communicated to the user. Further tasks, such as microplanning and surface realization, are included specifically into the NLG module.

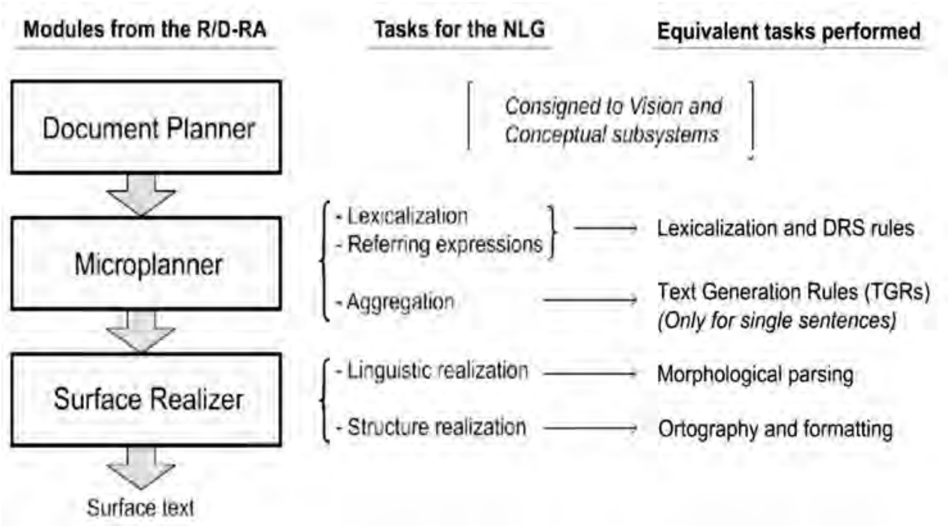


Figure 4 – *Schema of Reiter/Dale Reference Architecture (R/D-RA) [9], including the tasks related to each module that are necessary for a Natural Language Generator.*

The NLG module receives high-level semantic predicates from the reasoning stage, which are eventually converted into surface text, this is, a sequence of words, punctuation symbols, and mark-up annotations to be presented to the final user. There are several tasks to cover in order to carry out this process; they have been structured into the following stages:

1. Discourse Representation
2. Lexicalization
3. Surface Realization

Besides, the set of lemmata for the domain of interest has to be extracted from a restricted corpus of the specific language. The different corpora have been elaborated based upon the results of several psychophysical experiments on motion description, collected over a significative number of native speakers of the target language. In our case, ten different people have independently contributed to the corpus with their own descriptions of the sample videos, according to the capabilities of the tracking system. Four different languages have been implemented for this scenario: Catalan, English, Italian, and Spanish.

3.1. Representation of the Discourse

The chosen approach towards the implementation of semantics for NL generation is based on Discourse Representation Theory (Kamp & Reyle, 1993). This theory allows the construction of semantic structures representing linguistic information contained in NL sentences, in predicate logic formalism. Semantic relationships are stated by means of Discourse Representation Structures (DRSs). Here, the inverse process is implemented, consisting of the retrieval of NL text from logic predicates, by defining a set of DRS construction and transformation rules for each language.

DRSs are semantic containers which relate referenced conceptual information to linguistic constructions (Kamp & Reyle, 1993). A DRS always consists of a so-called universe of referents and a set of conditions, which can express characteristics of these referents, relations between them, or even more complex conditions including other DRSs in their definition. These structures contain linguistic data from units that may be larger than single sentences, since one of the ubiquitous characteristics of the DRSs is their semantic cohesiveness for an entire discourse.

One of the main semantic characteristics to take into account refers to cohesiveness. When a contextual basis is explicitly provided, the maintenance of the meaning for a discourse, including its cross-references, relations and cohesion can be granted. A particularly interesting and comprehensible example of discourse cohesion is the case of anaphoric pronominalization, which allows the generation of some referring expressions; for instance, we typically discard "*The pedestrian waits to cross. The pedestrian crosses*", in favor of "*The pedestrian waits to cross. S/he crosses*".

By using such structures, we will be able to point out the cross-references existing among the semantic constituents of a predicate. The classification of linguistically perceived reality into thematic roles (e.g. agent, object, location) is commonly used in contemporary linguistic related applications as a possibility for the representation of semantics, and justifies the use of computational linguistics for describing content extracted by vision processes. In the current implementation, these constituents can be classified as *agents*, *objects*, *locations*, and *events/situations*. Given that a situational analysis is accomplished for each detected agent, we take previously mentioned information about the focused agent as a basis to decide upon referenced expressions or full descriptions. An example which shows how the semantic representation and contextualization is undertaken by a DRS is illustrated in **Figure 5**. DRSs also facilitate the subsequent tasks for sentence generation. The syntactical features of a sentence are provided by the so-called Text Generation Rules (TGRs), which establish the position for the elements of the discourse within a sentence for a particular language. Due to the specific goals considered for this system, simple sentences are used for effective communication.

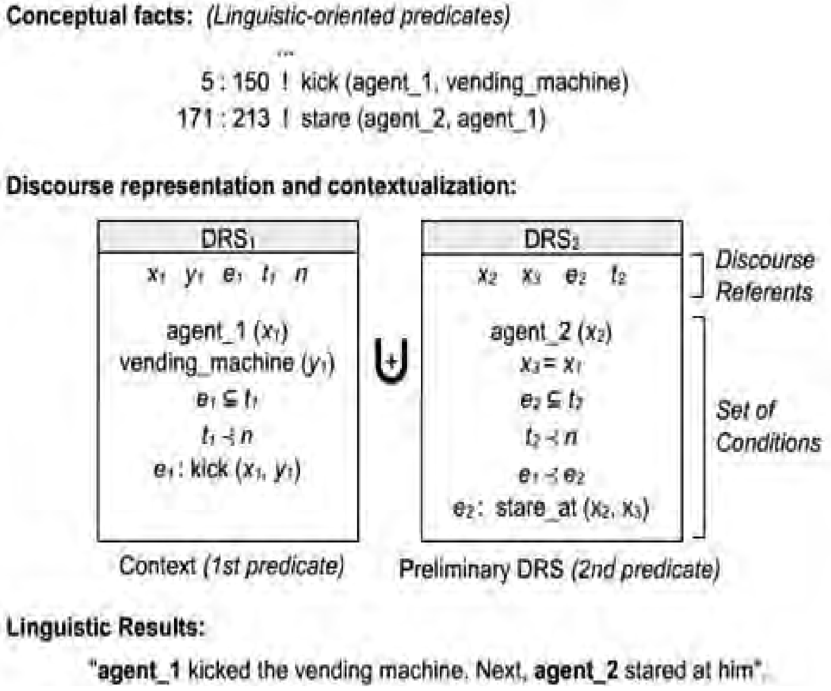

Conceptual facts: (Linguistic-oriented predicates)
...
5 : 150 ! kick (agent_1, vending_machine)
171 : 213 ! stare (agent_2, agent_1)
Discourse representation and contextualization:
DRS1
x1 y1 e1 t1 n
agent_1 (x1)
vending_machine (y1)
e1 ⊆ t1
t1 ≺ n
e1 : kick (x1, y1)
⊎
DRS2
x2 x3 e2 t2
agent_2 (x2)
x3 = x1
e2 ⊆ t2
t2 ≺ n
e1 ≺ e2
e2: stare_at (x2, x3)
Discourse Referents
Set of Conditions
Context (1st predicate)
Preliminary DRS (2nd predicate)
Linguistic Results:
"agent_1 kicked the vending machine. Next, agent_2 stared at him"

***Figure 5** – A pattern DRS allows the conversion of a stream of conceptual predicates into a string of textual symbols. Here, two predicates are validated. The first one instantiates a DRS, which serves as context for the following asserted facts. Once a new predicate is validated, it instantiates another DRS which merges with that context, thus providing a new context for subsequent facts. The temporal order of the events is stated by including them within time variables ($\varepsilon_1 \subseteq \tau_1$), placing these variables in the past ($\tau_1 \prec \nu$), and marking precedence ($\varepsilon_1 \prec \varepsilon_2$).*

The question of how to address temporal references also arises at the semantic level. A natural possibility consists of tensing the statement of recent observations in present perfect (e.g. *He has turned left*), and handling inferences in present tense (e.g. *He waits to cross*), although there exists a certain flexibility for the selection of tenses. A discourse referent for the utterance time of discourse (*n*) is required, so that the rest of temporal references t_i can be positioned with respect to it, see **Figure 5**.

3.2. Lexicalization

As stated in Reiter & Dale, 2000, *lexicalization* is the process of choosing words and syntactic structures to communicate the information in a document plan, i.e. the interpreted knowledge of logical predicates within a defined domain. Concretely, we will have to map the messages from the predicates, now linked by DRSs, into words and other linguistic resources that explain the semantic contents we want to communicate. It is difficult to bound the lexicalization process to a single module, since the mappings from semantic to linguistic terms are accomplished at several stages of the architecture; in this section we focus on lexicalization of prior knowledge, i.e. *agents*, *objects*, and *locations*, which have to be known beforehand.

The lexicalization step can be seen as a mapping process, in which the semantic concepts identifying different entities and events from the selected domain are attached to linguistic terms referring to those formal realities. This way, this step works as a real dictionary, providing the required lemmata that will be a basis for describing the results using natural language. Parsing processes will be in charge of traversing the syntactical structures obtained by the Text Generation Rules, and replacing the semantic identifiers by their suitable linguistic patterns. **Figure 6** shows an example of lexicalization for two aforeknown identifiers of semantic regions from the scenario.

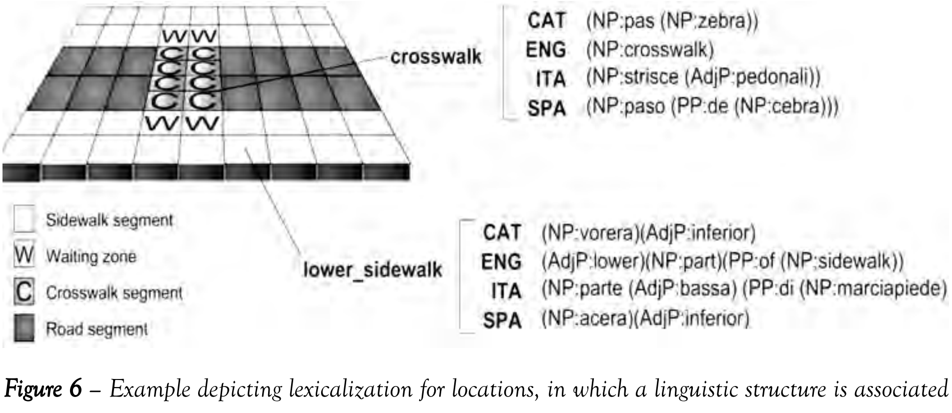

crosswalk
CAT (NP:pas (NP:zebra))
ENG (NP:crosswalk)
ITA (NP:strisce (AdjP:pedonali))
SPA (NP:paso (PP:de (NP:cebra)))
Sidewalk segment
W Waiting zone
C Crosswalk segment
Road segment
lower_sidewalk
CAT (NP:vorera)(AdjP:inferior)
ENG (AdjP:lower)(NP:part)(PP:of (NP:sidewalk))
ITA (NP:parte (AdjP:bassa) (PP:di (NP:marciapiede)
SPA (NP:acera)(AdjP:inferior)

Figure 6 *– Example depicting lexicalization for locations, in which a linguistic structure is associated with a semantic region of the scenario for each considered language. Only basic structural information is represented here, although morphological characteristics are also provided to the linguistic terms at this step.*

3.3. Surface Realization

The Surface Realization stage is accomplished in two steps. A first morphological process applies over each single word and partially disambiguates the individual abstraction of that word, by means of morphological attributions such as gender or number. These attributions can be propagated upon the semantic relations previously established by DRSs among the lemmata of a single piece of discourse. After that, a set of post-morphological rules was conceived to enable interactions among predefined configurations of words, thus affecting the final surface form of the text. This additional step is indispensable for many languages, in which certain phenomena force the surface form to change, e.g. contractions (*a* + *el* ➔ *al*, in Spanish), or order variation (*es* + *va* + *en* ➔ *se'n va*, in Catalan). **Table 1** shows some examples of morphological rules included in the grammar used for parsing.

1) $$\langle\text{"}go\text{"}\rangle \begin{pmatrix} verb \\ particip. \end{pmatrix} \xrightarrow{ENG} \langle\text{"}gone\text{"}\rangle\langle verb\rangle$$

2) $$\langle\text{"}meet\text{"}\rangle \begin{pmatrix} verb \\ particip. \end{pmatrix} \xrightarrow{ENG} \langle\text{"}met\text{"}\rangle\langle verb\rangle$$

3) $$\langle\alpha\rangle \begin{pmatrix} verb \\ particip. \end{pmatrix} \xrightarrow{ENG} \langle\alpha + \text{"}ed\text{"}\rangle\langle verb\rangle$$

4) $$\langle\text{"}a\text{"}\rangle\langle prep.\rangle + \langle\text{"}el\text{"}\rangle \begin{pmatrix} determ. \\ masc. \\ sing. \end{pmatrix} \xrightarrow{CAT\ ITA} \langle\text{"}al\text{"}\rangle \begin{pmatrix} prep. \\ determ. \\ masc. \\ sing. \end{pmatrix}$$

5) $$\langle\text{"}de\text{"}\rangle\langle prep.\rangle + \langle vowel + \alpha\rangle \xrightarrow{CAT\ ITA} \langle\text{"}d\text{'"}\rangle\langle prep.\rangle + \langle vowel + \alpha\rangle$$

6) $$\langle\alpha\rangle \begin{pmatrix} determ. \\ sing. \end{pmatrix} + \langle vowel + \beta\rangle \xrightarrow{CAT\ ITA} \langle\text{"}l\text{'"}\rangle\langle determ.\rangle + \langle vowel + \beta\rangle$$

Table 1 *- Examples of some simple morphological rules in Catalan, English, and Italian. Rules 1 and 2, in English, allow reducing the participle tag of a verb for two exceptions, and producing the word form. Rule 3 produces the participle in general. The other rules, for Catalan and Italian, deal with prosodic manipulation: rule 4 covers the contractions of a preposition with a determiner, and rules 5 and 6 are for apostrophication, when certain words appear in front of a word with an initial vowel..*

Finally, a general scheme for the entire process of generation is shown in **Figure 7**. The sentence "*He is waiting with another pedestrian*" is generated step by step from logical predicates, for the English language. The center column contains the tasks being performed, and the right column indicates the output obtained after each task.

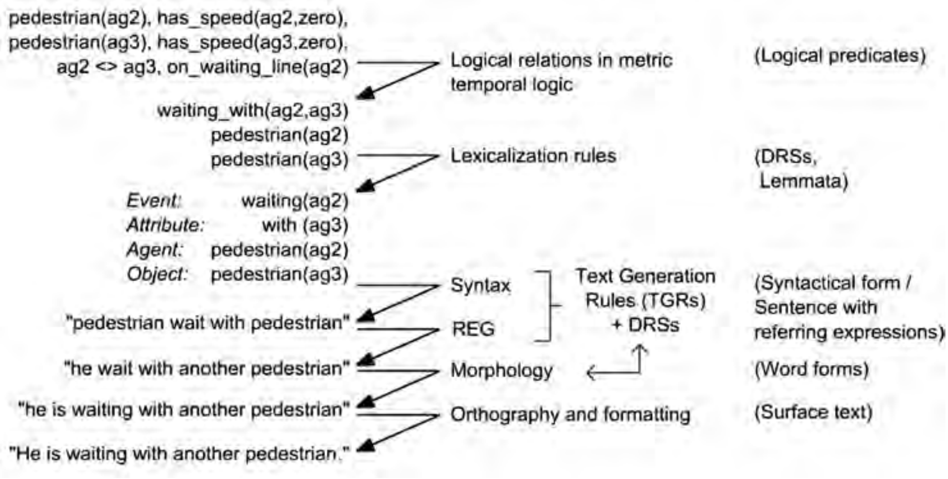
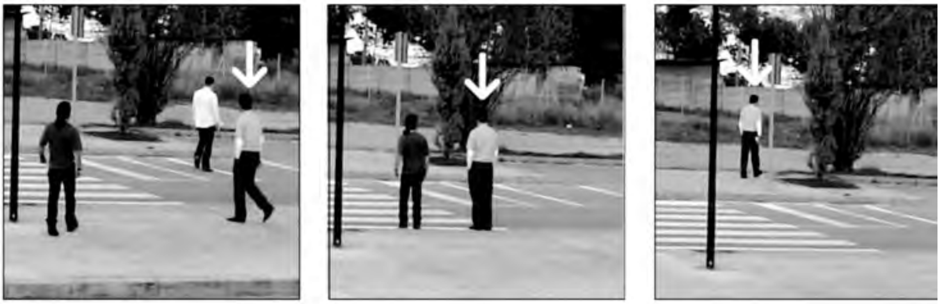

pedestrian(ag2), has_speed(ag2,zero),
pedestrian(ag3), has_speed(ag3,zero),
ag2 <> ag3, on_waiting_line(ag2)
Logical relations in metric temporal logic
(Logical predicates)
waiting_with(ag2,ag3)
pedestrian(ag2)
pedestrian(ag3)
Lexicalization rules
(DRSs, Lemmata)
Event: waiting(ag2)
Attribute: with (ag3)
Agent: pedestrian(ag2)
Object: pedestrian(ag3)
Syntax
REG
Text Generation Rules (TGRs) + DRSs
(Syntactical form / Sentence with referring expressions)
"pedestrian wait with pedestrian"
"he wait with another pedestrian"
Morphology
(Word forms)
"he is waiting with another pedestrian"
Orthography and formatting
(Surface text)
"He is waiting with another pedestrian."

***Figure 7** – Example for the generation of the sentence "He is waiting with another pedestrian" from logical predicates and for the English language.*

4. Experimental results

Next, some results are provided for the two scenes considered. For the first crosswalk scene, textual descriptions in Catalan, English, and Spanish have been selected for Agents 3 and 4, respectively. They include agents appearing or leaving, interactions with locations, and basic interpretations such as waiting with others to cross, or crossing in a dangerous way (i.e. crossing the road directly and not caring for vehicular traffic). The average measured time for generating a single sentence has been 3 milliseconds.



Pedestrian 3 (Catalan)

203 : *Lo vianant surt per la part inferior dreta.*

252 : *Va per la vorera inferior.*

401 : *S'espera per creuar.*

436 : *S'està esperant amb un altre vianant.*

506 : *Creua pel pas zebra.*

616 : *Va per la vorera superior.*

749 : *Se'n va per la part superior dreta.*

Pedestrian 3 (English)

203 : *The pedestrian shows up from the lower right side.*

252 : *S/he walks on the lower sidewalk.*

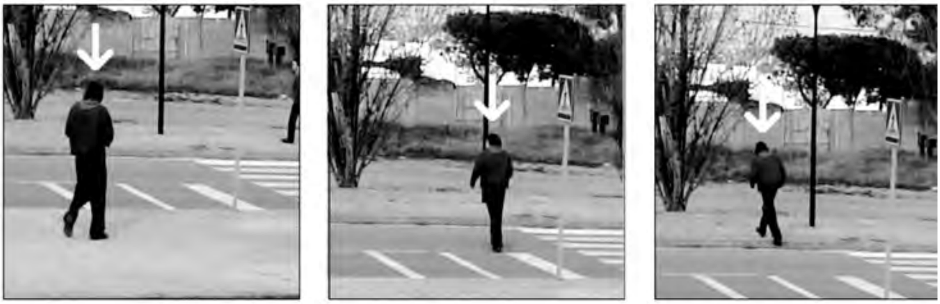
401 : *S/he waits to cross.*

436 : *S/he is waiting with another pedestrian.*

506 : *S/he enters the crosswalk.*

616 : *S/he walks on the upper sidewalk.*

749 : *S/he leaves by the upper right side.*



Pedestrian 4 (Spanish)

523 : *El peatón aparece por la parte inferior izquierda.*

572 : *Camina por la acera inferior.*

596 : *Cruza sin cuidado por la calzada.*

681 : *Camina por la acera superior.*

711 : *Se va por la parte superior izquierda.*

Pedestrian 4 (English)

523 : *The pedestrian shows up from the lower left side.*

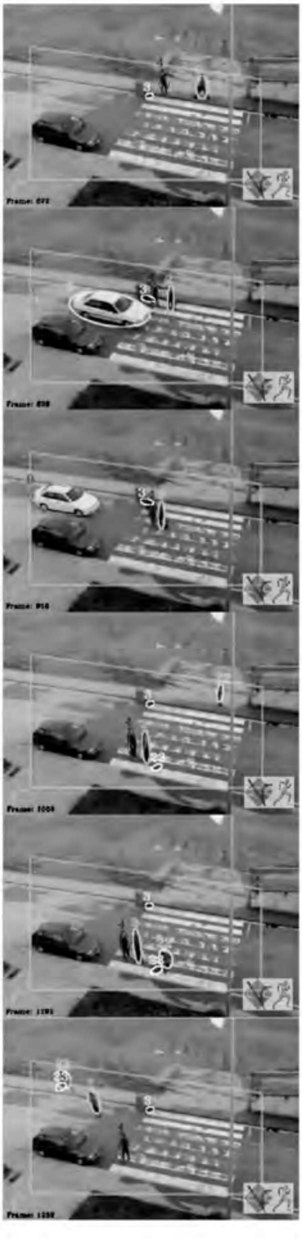
572 : *S/he walks on the lower sidewalk.*

596 : *S/he crosses the road carelessly.*

681 : *S/he walks on the upper sidewalk.*

711 : *S/he leaves by the upper left side.*

Some results for the second scene are presented in Catalan, Italian, and English. In this case there exist more complex interactions and interpretations of events, e.g. abandoned objects, dangers of run over, thefts, or chasings.



470 : *Un vianant surt per la part superior asquerra.*
470 : *A pedestrian appears from the upper left side.*
470 : *Un pedone compare nella parte superiore sinistra.*

492 : *Lo vianant camina per la vorera superior.*
492 : *Il pedone cammina sulla parte alta del marciapiede.*
492 : *The pedestrian walks on the upper part of the sidewalk.*

583 : *Gira pac a la dreta per la part superior de lo pas zebra.*
583 : *S/he turns right in the upper part of the crosswalk.*
583 : *Gira a destra sulla parte alta delle strisce pedonali.*

591 : *S'ha parat allà mateix.*
591 : *S/he has stopped in the same place.*
591 : *Si è fermato in questa posizione.*

615 : *Ha dixat l'objecte a terra.*
615 : *S/he has left an object.*
615 : *Ha lasciato un oggetto in terra.*

630 : *Un nou vianant surt per la part superior dreta.*
630 : *A new pedestrian appears from the upper right side.*
630 : *Un altro pedone compare nella parte superiore destra.*

642 : *Lo vianant camina per la vorera superior.*
642 : *The pedestrian walks on the upper part of the sidewalk.*
642 : *Il pedone cammina sulla parte alta del marciapiede.*

656 : *Lo primer vianant camina per allà mateix.*
656 : *The first pedestrian walks on the same place.*
656 : *Il primo pedone cammina in questa zona.*

687 : *L'objecte pareix haver astat dixat a la part superior de lo pas zebra.*
687 : *The object seems to have been abandoned in the upper part of the crosswalk.*
687 : *L'oggetto sembra che sia stato abbandonato nella parte alta delle strisce pedonali.*

692 : *Lo primer vianant s'ha trobat en lo segon vianant allà mateix.*
692 : *The first pedestrian has met the second pedestrian in the same place.*
692 : *Il primo pedone si è incontrato con il secondo pedone in questa posizione.*

822 : *Un vehicle pareix que astà a punt d'atropellar lo primer vianant.*
822 : *A danger of runover between the first pedestrian and a vehicle seems to have been detected.*
822 : *Un veicolo ha rischiato d'investire il primo pedone.*

4.1. Evaluation

Thirty English native speakers were recruited from among different sources in 5 countries. Less than one third of the subjects are members of a computer science department, and none of them has NLP background. Subjects were told to describe both sequences in a natural and linguistically correct manner, using the expressions they considered most suitable. Only the results concerning the second sequence (the one showing vehicular traffic) are presented, since they contain a greater number of facts and also enable more subjective interpretation, and thus state the main problems of our approach better.

The ground truth of the second sequence contains 30 basic facts, and thus a limitation has been imposed to make subjects describe a number of facts comprised between 20 and 40, in order to deal with comparable levels of detail. However, they were free to include them into either single or compound sentences; e.g. "a kid runs, then falls down, and finally gets up again" was considered as 3 events in one sentence. **Figure 8** presents statistical graphs concerning the population and the basic results of the experiment.

4.1.1. Qualitative results

The qualitative comparison between the generated data and the collected set has been done regarding several concerns: the main objective at a semantic level has been to detect the differences between the set of facts detected by the subjects and the one generated by the system. On the other hand, we also wanted to learn the mechanisms of reference used, and which kind of words, expressions, and connectors were being most employed. These were compared to our choices. When considering the list of facts to compare to the inputs, facts having closely related meanings have been gathered together, e.g., "*notice*"-"*realize*", or "*run after*"-"*chase*"-"*chase after*".

- A practical rule for *simplicity* is deduced from the results. The majority of cases avoid obvious explanations that can be logically derived from a more expressive linguistic choice. When one states "*A man walks down the sidewalk*", there is no need to include "*A man appears*". Also, there is no need to state that a person is "bending" when picking up an object; it is obvious when the object is known to be on the ground.

- The greater difference regarding expressiveness happens when the subjects deduce the *intentions* of the agents by the context, using common sense. For instance, *"He waves his hands in amazement that the car didn't stop*" or "*He seemed somewhat hesitant*". Sometimes, the following situations in the scene are anticipated, like "*A person is walking to the zebra crossing to meet someone*". These constructions are very useful to conduct the discourse.

- One of the main tasks lacking in the described generation system is the *aggregation* of simple sentences into more complex and expressive ones, using mechanisms such as coordination or subordination. This has the main advantage of emphasizing certain elements of the discourse. For instance, *"After crossing the street they stop, one of them puts his bag on the ground, and while they are talking, another guy comes and snatches the bag*" prioritizes the object left over the crossing and the theft over the talk.

- The use of certain adverbs, adjectives, and other complementary words has been seen as helpful towards a richer discourse: "*nearly* hit by an *oncoming* vehicle", "jumps back *in surprise*", "move back *slightly*", "they *only* crossed the street half-way", among others.

4.1.2. Quantitative results

In order to retrieve some quantitative measures about the adequacy of the proposed generation, in the following we provide some statistical results that compare the frequencies of use of the two sets of facts, the generated and the collected ones. The main purpose is to decide up to which point the current capabilities of the vision tracking and conceptual reasoning levels enable us to provide natural results.

Table 2 shows the whole list of facts, containing those used for generation and those detected by the subjects of the experiment, sorted by frequency of use. Based on these sorted results, we easily identify facts that should be included, replaced, or removed from the current set. The list of descriptions done by the system contains many of the most frequent facts: the system generates 100% of the facts used by more than half of the participants, and 77.8% of the ones employed above the average share of facts (25.9%). The average share has been computed as the average of the proportions of the facts in the whole list.

- First, we notice some of them referring to the same situations in different manners, like "danger of run over" and "almost hit / knock down / run over pedestrians"-"pass without stopping / not let pedestrians cross". Selecting suitable terms depends on the purposed application, and hence it is not identified as a main concern.
- Concerning facts to add, the most significative ones are those referring to the talk and interactions between the first two pedestrians ("talk", "shake hands", "greet each other", "wave", "chat"). This points out an actual limitation of the tracking system, which cannot provide such detailed information about body postures at the moment.
- Finally, some of the facts used should be discarded. Some features have been detected which seem to indicate that facts are not interesting enough to be included in the final text, such as being obvious (*reach the other side after crossing*, *bend to take an object*), being an uncommon expression (*having an exchange*, *motioning someone somewhere*), being too subjective (*two people being friends*), or guessing emotions (*to seem confused, angry, or surprised*). When a situation can be interpreted in more than one way, each of the interpretations gets less support than a unique one, so that uncertainty is another factor to consider.

In addition, it is also interesting to notice that just about one quarter of the population has included *color references* to support their descriptions. Most of these (above 70% of them) use a single reference, for the "white car", which is the only agent with a very distinctive color.

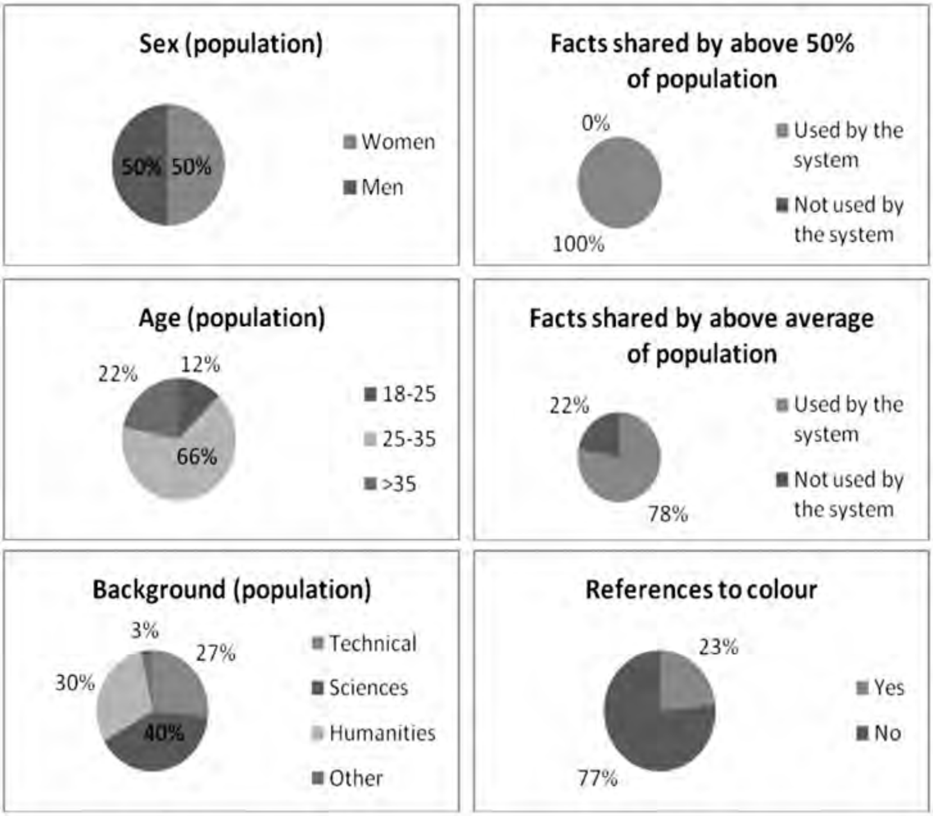


Figure 8 *– Statistics about the NL generation experiment, for English and the outdoor sequence. The population consisted of 30 subjects from different backgrounds. The left column contains information about the population, the right column shows quantitative results about the evaluation and comparison with the facts used.*

List of detected facts, sorted by frequency of use

% Used	Fact
100%	**Ped1 leaves object1**
100%	**Peds1,2 cross / try to cross / walk to other side / want to cross**
90.0%	**Ped1 walks**
86.7%	**Ped2 leaves Obj2**
83.3%	**Ped3 runs / runs off / runs away**
83.3%	**Peds1,2 enter_crosswalk / cross / go across / go on crossing**
83.3%	**Veh2 gives way / stops / wait for them to cross**
80.0%	**Ped2 chases / chases after / runs after Ped3**
70.0%	**Ped3 picks up / grabs / snatches Obj2**
63.3%	**Peds1,2 meet / stand close**
60.0%	**Ped3 appears / enters**
50.0%	**Ped3 crosses**
50.0%	**Ped3 steals / thief**
50.0%	**Peds2 walks / comes**
46.7%	**Ped3 walk / approaches /comes**
46.7%	Veh1 passes without stopping / not allowing them to cross
46.7%	**Veh2 appears / comes**
43.3%	**Peds1,2 back up and stop / pull back**
43.3%	Peds1,2 talk / chat / converse (1st time)
40.0%	**Ped1 stops / reaches crosswalk**
40.0%	**Ped2 appears**
40.0%	**Peds1,2 stop / stand (2nd time)**
40.0%	**Veh1 appears / comes**
36.7%	Peds1,2 notice/realize/see Ped3
36.7%	Veh1 almost hits / almost knocks down / almost runs over Peds1,2
33.3%	Ped2,3 run
33.3%	Peds1,2 shake hands (1st time)
26.7%	Ped1 holds briefcase / ...with a bag
26.7%	Peds1,2 greet each other
26.7%	Peds1,2 talk/converse/chat (2nd time)
23.3%	**Ped1 appears**
20.0%	Ped1,2 keep on talking / while they talk (while crossing)
20.0%	**Peds1,2 stop at Veh1**

20.0%	Veh2 arrives / approaches the zebra
16.7%	**object1 abandoned / forgotten**
13.3%	Ped2 waves / attracts attention of Ped1
13.3%	Peds1,2 shake hands (2nd time)
13.3%	Peds1,2 still talking / keep on chatting (2nd time)
13.3%	**Peds2,3 leave**
13.3%	Veh1 accelerates / goes on
13.3%	Veh1 reaches / runs towards / approaches
13.3%	**Veh2 exits / passes by**
10.0%	**danger of run over / about to run over**
10.0%	Ped1 eventually follows the chase
10.0%	Ped1 stays watching
10.0%	Ped1,2 start talking (2nd time)
10.0%	Ped3 does not notice / ignores Obj1
10.0%	Ped3 walks away from them
10.0%	shout at the driver
10.0%	**Veh2 accelerate /drives on**
6.7%	Ped1 says hello to Ped2
6.7%	Ped1 spins around confused / looks on bewildered / seems hesitant
6.7%	Ped1 walks away
6.7%	Ped2 reaches / arrives to Ped1
6.7%	Ped2 tries to recover/reclaims his bag
6.7%	Peds1,2 complain against / protest to car driver / raise-wave hands
6.7%	Peds1,2 dont notice Ped3
6.7%	Peds1,2 dont pay attention when crossing
6.7%	Peds1,2 reach the other side
6.7%	Peds1,2 say goodbye to each other
6.7%	Peds1,2 wait to let Veh2 pass
6.7%	**Veh1 leaves**
3.3%	brief exchange between Peds1,2
3.3%	Ped1 checks road
3.3%	Ped1 motions Ped2 to cross
3.3%	Ped1 motions Ped2 to cross
3.3%	Ped1,2 have a brief exchange
3.3%	Ped1,2 out of range of vehicles
3.3%	Ped2 tells Ped1 about Ped3

3.3%	Ped3 bends down
3.3%	Ped3 ducks
3.3%	Ped3 notices Obj2
3.3%	Ped3 stops near Obj2
3.3%	Peds 1,2 seem to be friends
3.3%	Peds1,2 are angry at Veh1
3.3%	Peds1,2 are surprised
3.3%	Peds1,2 communicate
3.3%	Peds1,2 let the car continue its way
3.3%	Peds1,2 wait for car to pass
3.3%	**Veh1 brakes up**

Table 2 - Percentages of use for the facts detected by the users. Facts are described in a schematic way (Ped=Pedestrian, Veh=Vehicle, Obj=Object). Shadowed facts are currently being used for automatic generation. Percentages above average use have been colored green; the ones below 10% have been colored red.

5. Conclusion

From the qualitative considerations, we notice a great difficulty: how to balance objectively certain facts with more expressive but also more uncertain facts in order to obtain a consistent description of what is happening, by also making the generated text relevant and communicative. The system should be enhanced with a richer use of complementary words, and on the other hand, simple sentences should be aggregated as needed to lead the discourse to defined goals. Some behavioral models should be included, too, in order to introduce interpretations regarding intentionality of the agents.

The quantitative results obtained provide objective information about the facts that should be considered. Nevertheless, some of the less frequently used facts are very appropriate for the description, such as people "*having brief exchanges*" or a pedestrian "*motioning*" someone somewhere. This suggests running new experiments, where the subjects could choose among different expressions to refer to the situations observed. In this way, the subjects would not be limited by not finding a suitable expression.

Regarding multilinguality, the system performed in the same way and presented the same problems in English as in the rest of the languages considered.

Most of the limitations for the described NLG module come clearly determined by the restrictive domain of work. The linguistic models need to be extended as new situations can be detected by the HSE system, since the content to be communicated is provided entirely by the situational analysis. The deterministic approach that has been chosen limits the variety of produced sentences, but ensures that the output results will be linguistically correct, since they obey the constructions proposed by native speakers and encoded into the models.

The modular architecture proposed for the NLG subsystem apparently allows the common stages to remain unchanged, disregarding the incorporation of new languages or the enlargement of the detection scope. So far, the addition of a new language has only required extending DRS rules and parsing grammars, which allows for a fast and effective implementation of similar languages.

Further steps include an enhancement of the Microplanner to support sentence aggregation. This would allow ordering the information structured in single sentences and mapping it into more complex sentences and paragraphs. Discourse Representation Theory has been proved consistent to accomplish this task (Kamp & Reyle, 1993).

6. Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). Jordi Gonzàlez also acknowledges the support of a Juan de la Cierva Postdoctoral fellowship from the Spanish MEC.

7. References

Arens, M., & Nagel, H.-H. 2003. "Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences". *Proc. of the KI 2003. 1,* pp. 149-163. Berlin, Heidelberg, New York: Springer-Verlag.

Buxton, H., & Gong, S. 1995. "Visual surveillance in a dynamic and uncertain world". *AI-Magazine,* I (78), 431-459.

Emele, M., Heid, U., Momma, S., and Zajac, R. 1990. "Organizing linguistic knowledge for multilingual generation", *Proceedings of the 13th conference on Computational linguistics* Volume 3, pages 102-107. Association for Computational Linguistics Morristown, NJ, USA.

Fernández, C., Baiget, P., Roca, X., & Gonzàlez, J. 2007. "Semantic Annotation of Complex Human Scenes for Multimedia Surveillance". AI*IA 2007. *Tenth Congress of the Italian Association for Artificial Intelligence.* (pp. 698-709). Roma, Italy: Springer LNAI.

Gerber, R., & Nagel, H.-H. 2008. "Representation of Occurrences for Road Vehicle Traffic". (Elsevier, Ed.) *Artificial Intelligence,* 4-5 (172): 351-391.

Gonzàlez, J., Rowe, D., Varona, J., & Roca, F. 2008. "Understanding Dynamic Scenes based on Human Sequence Evaluation". *Image and Vision Computing.* 10.1016/j.imavis.2008.2.004.

Hovy, E.H. 1993. "Automated discourse generation using discourse structure relations". *Artificial Intelligence,* 63 (1-2): 341-385

Huerta, I., Rowe, D., Mozerov, M., & Gonzàlez, J. 2007. "Improving Background Subtraction based on a Casuistry of Colour-Motion Segmentation Problems". *3rd IbPRIA:* 475-482. Girona, Spain: Springer LNCS.

Kamp, H., & Reyle, U. 1993. *From Discourse to Logic.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kojima, A., Izumi, M., Tamura, T., and Fukunaga, K. 2000. "Generating Natural Language Description of Human Behavior from Video Images", ICPR-2000, vol. 4:728-731.

Kojima, A., Tamura, T., & Fukunaga, K. 2002. "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions". *International Journal of Computer Vision,* II , 50:171-184.

Lou, J., Liu, Q., Tan, T., and Hu, W. 2002. "Semantic interpretation of object activities in a surveillance system". *Proc. Int. Conf. Pattern Recognition,* 777-780.

Marburger, H., Neumann, B., & Novak, H. 1981. "Natural Language Dialogue about Moving Objects in an Automatically Analyzed Traffic Scene". *Proc. IJCAI-81, Vancouver.*

Nagel, H.-H. 2004. “Steps toward a Cognitive Vision System”. *AI Magazine*, II (25): 31-50.

Nevatia, R., Zhao, T., and Hongeng, S. 2003. “Hierarchical language-based representation of events in video streams”. *Proc. IEEE Workshop on Event Mining.*

Reiter, E., & Dale, R. 2000. *Building Natural Language Generation Systems.* Cambridge: Cambridge University Press.

Rowe, D., Rius, I., Gonzàlez, J., & Villanueva, J. 2005. “Improving Tracking by Handling Occlusions”. *3rd ICAPR:* 384-393. UK: Springer LNCS.

Wilson, R., & Keil, F. (eds.). 2001. *The MIT Encyclopedia of the Cognitive Sciences.* Cambridge: MIT Press.

NATURE OF THE ARTICLES

Computational Linguistics, Foreign Language Teaching and Learning, Forensic Linguistics, Language for Specific Purposes, Language Planning, Second Language Acquisition, Speech Pathologies, Translation.

FORMAT OF THE ARTICLES

Contributions should be written in English, using the software package Word. Title of the paper and name, address, telephone number and e-mail address of the author (s) should be included on a separate sheet. Submissions must be sent by e-mail attachment.
For information about length, abstract, references, etc. please use the web page www.webs.uvigo.es/vialjournal/

All correspondance should be addressed to:
Rosa Alonso ralonso@uvigo.es or Marta Dahlgren dahlgren@uvigo.es